Indirect evidence of reporting biases was found in a survey of medical research studies

Loai N. Albarqouni[1], José A. López-López[2], Julian PT Higgins[2]

1. Centre for Research in Evidence-Based Practice (CREBP), Faculty of Health science and Medicine, Bond University, Gold Coast, Australia.
2. School of Social and Community Medicine, University of Bristol, Bristol, UK

Corresponding author

Loai Albarqouni, MD, MSc.
Centre for Research in Evidence-Based Practice, Faculty of Health Sciences and Medicine,
Bond University, QLD, Australia 4226.
Email: lnb6des@daad-alumni.de

Abstract

Objective: To explore indirect evidence of reporting biases by examining the distribution of P-values/z-scores reported in published medical articles, and to compare P-values/z-scores distributions across different contexts.

Methods: We selected a random sample (n=1500) of articles published in PubMed in March 2014, and included articles that reported sufficient details of the results of inferential statistics. Additionally, we extracted information on study type, design, medical discipline and P-values/z-scores for the first-reported outcome and primary outcome (if specified) from each article.

Results: Out of the 1500 randomly selected records, 758 (50.5%) were included. We retrieved or calculated 758 P-values/z-scores for first-reported outcomes and 389 for primary outcomes (specified in only 51% of included studies). The first-reported and the primary outcome differed in 28% (110/389) of the included studies. The distributions of P-values/z-scores for first-reported outcomes and primary outcomes showed a notable discontinuity at the common threshold of statistical significance (P-value=0.05/z-score=1.96). A caliper test showed an imbalance in the z-scores around the common significance threshold using 5% and 10% caliper sizes for the first reported outcomes as well as primary outcomes. We also found marked discontinuities in the distributions of z-scores across various medical discplines, study designs and types.

Conclusions: Reporting biases are still common in medical research. We discuss its implications, strategies to detect it and recommended practices to avoid them.

Keywords: bias, p-curve, p-hacking, methodology, reporting bias, publication bias

Work count: Abstract: 219, Main text: 2944, Tables: 2, Figures: 4, References: 40, Appendices: 2.

What is new?

Key finding

- There are discontinuities of the distribution of p-values/z-values at the typical thresholds of statistical significance that may provide indirect insights on reporting bias
- Similar results were observed across various study designs and types.

What this study adds to what is known?

- Notable peaks in the distributions at common thresholds of statistical significant are consistent with either suppression of non-statistically significant results or 'manipulation' of reported findings to reach statistical significance.
- The outcome that is reported earliest in an article is more prone to this phenomenon than the primary outcome.

What is the implication, and what should change now?

- The present investigation underpin the importance of the efforts and initiatives to tackle the mechanisms causing reporting biases (e.g. registration of studies, protocols and statistical analysis).
- Researchers should continue to be encouraged to emphasize confidence intervals and effect sizes, rather than P-values, in the interpretation of results.
- There is a need for advocating the importance of replication, as well as the benefits of complete publication of research findings to reduce the prevalence of reporting biases in scientific literature

Introduction

Complete publication of study results is essential to allow healthcare professionals and policy makers to make informed decisions. However, selective or distorted reporting is frequent in medical research.[1] Reporting biases arise if dissemination of research findings is influenced by the nature of the results. If undetected, reporting biases can lead to inaccurate conclusions and inappropriate decisions about health care and resource allocation, with potentially serious implications.[2] Failure to publish research findings honestly is unethical and a form of research misconduct.[3, 4] Furthermore, research inaccessibility leads to waste of limited resources, unnecessary duplication, and loss of trust in scientific integrity.[5]

Reporting biases may impact scientific reports in different ways.[6-9] First, a whole study may be suppressed, or harder to find, or published with delay, if its results are not considered to be interesting. The label 'publication bias' is typically used to refer to this phenomenon.[10] Publication bias is the form of reporting bias that has been most extensively discussed in the literature over the last 60 years. [11-13] Second, results within a report of a study may be biased if the authors report the most interesting findings. For example, they may report the finding with smallest P-value or largest effect estimate after performing several analyses on the same outcome. Several terms have been coined to refer to such practice, including selective analysis reporting, data dredging and p-hacking.[14] Alternatively, some outcomes that were measured and analysed may be missing if the authors did not consider the results to be interesting.

Although these reporting biases are likely to have been always present in the dissemination of research findings, more attention has been drawn to them recently due to the widespread use of systematic reviews. The validity of conclusions drawn from systematic reviews, intended to summarize the state of the art in a scientific area, is threatened if published results are not representative of the population of all conducted studies and analyses. Meta-analysis provides researchers with several graphical methods and statistical tests to assess the possible presence of reporting biases.[6, 10, 13, 15] The exponential growth of published meta-analyses, many of them including some assessment of reporting biases, is likely to have increased the concern of incomplete publication of results as an ubiquitous problem in the scientific literature.[8]

Evidence of reporting biases can be direct or indirect. Direct evidence includes tracking of cohorts of registered studies or conference proceeding abstracts and comparing the results of published and unpublished findings. For instance, studies have provided empirical evidence that studies with significant or positive results were more likely to be published, or more likely to be published earlier, than those with non-significant or unimportant results. [5, 8] Direct evidence may also come from the acknowledgement of bias by those involved in the publication process, such as researchers, referees and editors.[16]

Indirect sources of evidence of reporting biases include the observation of a disproportionately high percentage of statistically significant findings in the published literature, as well as notable discontinuities in the P-value/z-score distribution curve just above the main significance thresholds (p=0.05/z-score=1.96). Several papers have been published illustrating similar approaches in psychology, sociology and natural science.[14, 17-19] Here we aim to explore indirect evidence of reporting biases by examining the empirical distribution of P-values/z-scores reported in a large set of medical research studies, and to compare this distribution across different contexts.

Methods

Study eligibility and selection

We conducted a descriptive cross-sectional survey of peer-reviewed, published, medical research articles. We sought original, primary and quantitative research articles, and searched the PubMed database using a simple search strategy that would identify most of these (Appendix 1). We restricted the search to articles published in March 2014, and selected a random sample of 1500 of the identified articles. To be included in the analyses, articles had to be written in English and had to involve only human participants. Articles had to include inferential statistics that investigated the efficacy or side effects of a medical or surgical intervention; or investigating risk factors, exposures or prognostic factors (epidemiological associations). We considered a wide range of study designs including randomized clinical trials, controlled clinical trials, before-after trials, cohort studies, case-control studies and cross-sectional studies, and we considered a wide range of estimates including differences in means, risk ratios, odds ratios, hazard ratios, correlations and regression coefficients. We included only articles that either reported the P-value or provided sufficient information to calculate a P-value for either the first reported or the primary outcome. We excluded duplicate reports of the same study as well as inaccessible full-text articles (e.g. published abstracts without full articles, or study protocols).

Data screening and extraction

We developed a standardised data extraction form, which was pilot-tested by all members of the research team. We extracted data based on the first reported outcome in the abstract (preferentially) or in the results section. For each included article, we extracted the following information:

- Author list and citation details.
- Medical speciality: we used the categories suggested by Davey et al.[20]
- Study type: therapeutic/intervention, prognostic, aetiological/risk factor.
- Study design: we used the classification used by Grimes et al:[21] randomized controlled trial (RCT), non-RCT, cohort, case-control, cross-sectional.
- Sample size: total sample size used in the analysis which yielded the P-value.

- Whether the primary outcome was specified (Yes/No) and whether it was the same as the first reported outcome (Yes/No/Unclear).
- 2-sided P-value, or information sufficient to calculate it, for the first reported outcome and for the primary outcome (if specified). We used the following hierarchy to determine each P-value, where only one of the following types of information was required:
    1. Exact 2-sided P-value: from the hypothesis test.
    2. Effect estimate with standard error or confidence limits: We used methods described by Altman and Bland to calculate P-values[22] from these measures.
    3. Test statistics: Z, chi-squared, t or F statistic, with degrees of freedom if applicable.
    4. For two-group designs reporting continuous outcome data: sample size, mean and standard deviation (or standard error) for each group.
    5. For studies reporting dichotomous outcome data: contingency table (e.g. 2×2 table).

Where a specified primary outcome differed from the first reported one, we implemented the same hierarchy to extract a P-value for each of the two outcomes.

Data analysis

As a first step, we transformed the two-sided p-values into z-scores, and used the latter as our main dependent variable. We plotted the distribution of z-scores across all included studied, both for first reported outcomes and for primary outcomes, using histograms. In the absence of any bias and if all effects are truly null, these z-scores would be uniformly distributed. We repeated these plots with subsets of the studies to explore the distributions of z-scores stratified by medical specialty, study design and study type. Moreover, we used the caliper test described by Gerber and Malhotra to explore the existence of discontinuities in the distribution of z-scores around the critical value of 1.96[19]. With regards to p-values, we compared the frequency of values in equal sized intervals just below and just above the threshold values commonly used for statistical significance (0.01 and 0.05), using a chi-squared test. We performed all analyses using the R statistical software (version 3.2.3).[23]

Results

Description of included studies

Figure 1 displays the study selection process in a flow chart. Of the 1500 randomly selected articles, we included 758 (50.5%). Among these included articles, 422 (56%) described therapeutic/intervention studies, 207 (27%) were aetiological/risk factor studies, and 129 (17%) were prognostic studies. With regards to study design, 264 (35%) were RCTs, 53 (7%) were non-RCTs, 145 (19%) were cross-sectional, 238 (32%) were cohort and 55 (7%) were case-control studies.

The medical disciplines of the included articles were cancer (105; 14%), cardiovascular (116; 15%), central nervous system/musculoskeletal (97; 13%), digestive, endocrine, nutritional and metabolic (98; 13%), gynaecology/pregnancy/birth (58; 8%), infectious (44; 6%), mental health/behavioural (75; 10%), urogenital (33; 4%), respiratory (21; 3%) and other disorders (111; 14%). The sample size of all included studies ranged from 6 to 375,888, with a median of 142 participants (range 55-525; IQR=470).

Out of the 264 included RCTs, the primary outcome was specified in 190 (72%). The primary outcome was also the first reported outcome in 143 (75%) studies, while it was not the first reported outcome in 45 (24%) and unclear in 2 (1%). In studies other than RCTs, the primary outcome was specified only in 199 out of 494 included studies (40%). The primary outcome was also the first reported outcome in 133 (67%) studies, while it was not the first reported outcome in 65 (33%) and unclear in one study.

The 742 excluded articles comprised 245 (33%) with only descriptive statistics, 144 (19%) with no original data, 121 (16%) with inaccessible full texts, 84 (11%) that were diagnostic or cost-analysis studies, 58 (8%) without sufficient information to extract or calculate a P-value, 48 (7%) with non-human research participants and 42 (6%) that were qualitative.

Empirical distribution of z-scores and p-values

We retrieved 758 results for first reported outcomes, with a median P-value of 0.011[0.0006 - 0.45] (z-score: 2.29[3.24-0.126]). Figure 2 shows the distribution of z-scores for first reported outcomes and primary outcomes, with dashed vertical lines for the common threshold of p = 0.05/z = 1.96 for statistical significance. In both distributions, there is a clear majority of z-scores above 1.96. Of particular note is the dramatic spike in the frequency of z-scores just

over the significance threshold z-score of 1.96 (P-value = 0.05). The results of the caliper tests using 5% and 10% caliper sizes for the first reported outcomes as well as primary outcomes showed a notable imbalance in the numbers of findings around the common significance threshold of 1.96 (P-value = 0.05), which is evident across the two caliper sizes (5% and 10%) in both first reported outcomes and primary outcomes (Table 1).

Table 2 shows that the majority of the retrieved P-values (both for first reported and primary outcomes) were smaller than common significance thresholds (0.05 and 0.01) with a total of 592 (78%) P-values of first reported outcomes were equal to or smaller than 0.05, and 376 (50%) were also equal to or smaller than 0.01.

The distribution of P-values reported in included studies for the first reported outcomes compared with the primary outcomes and grouped by study design (RCTs vs. studies other than RCTs). It shows that P-values more likely to be significant for the first reported outcomes than for the primary outcomes in RCTs only (*p*-value = 0.02391) (Table 3).

Stratified analyses

Figure 3 shows the histograms of z-scores for first reported outcomes stratified by medical speciality, annotated by median sample sizes within specialties. All figures reflect the same pattern of a majority of z-scores over the threshold of statistical significance, but the distributions appear less skewed in some of the disciplines with larger average sample sizes, namely infectious diseases (n=28; 63.6% of z-scores above 1.96), urogenital (n=21; 63.6%) and cancer (n=75; 71.4%). The most extreme patterns appeared in the area of respiratory diseases (n=19; 90.5% of z-scores above 1.96), cardiovascular (n=97; 83.6%) and central nervous system or musculoskeletal disorders (n= 78; 80.4%). Similar trends were observed in the histograms of z-scores for primary outcomes stratified by medical speciality, which are provided in Appendix 2.

Histograms of z-scores for first reported outcomes did not show major differences in the distribution according to study design (Figure 4). Likewise, we obtained similar histograms when exploring the distributions of z-scores for first reported outcomes stratified by study type, and also when plotting the distributions of z-scores for primary outcomes stratified by study design or by study type (Appendix 2).

Discussion

Our distributions of reported P-values/z-scores from medical research studies show notable peaks (or discontinuities) in the distributions at the common threshold of statistical significance (z-score = 1.96/p = 0.05) that may provide indirect insights on reporting bias. The outcome that is reported earliest in an article is more prone to this phenomenon than the primary outcome. Only about half of the included articles specified the primary outcome, and in 28% of the articles the first reported outcome was not the primary outcome. Similar patterns were observed across various medical discplines, study designs and types.

Strengths of our study include use of a large random sample of 1500 articles recently published, of which 758 contributed to the analysis. We also implemented manual data extraction from the articles; provided a breakdown by medical disciplines, study type and study design; and computed z-scores/P-values when they were not reported directly. However, we were unable to retrieve all of the articles listed in our random sample (see Figure 1). We are unable to draw any conclusions about whether the observed distribution is due to data manipulation ('p-hacking') or genuine effects, because as Bruns and Ioannidis suggested[24], p-curves may neither identify genuine effects nor p-hacking in observational research.

The presence of reporting biases has been claimed repeatedly in the medical literature,[1, 4, 8] and in other areas as diverse as cognitive sciences,[17, 25] biology,[26] educational research,[27] political sciences,[28] and management research.[7] Although definitions of reporting biases and strategies to explore vary, the conclusions and implications for researchers are similar across disciplines. Previous studies have investigated empirical distributions of published P-values. A study of abstracts in PubMed reported an extremely skewed distribution of P-values, with a substantially higher proportion of P-values below 0.05 in non-randomized studies compared to randomized trials.[29] In a review of meta-analyses, Ioannidis and Trikalinos also concluded that significant P-values were overrepresented [30]. In psychology, some studies also explored the P-value distribution and showed an inordinately high number of P-values just below 0.05.[17, 31]

It is good practice to specify the primary outcome before performing the statistical analysis of a clinical trial.[9] In our survey, we found that 72% of the RCTs vs. 40% of studies other than RCTs specified the primary outcome in their reports. Moreover, the specified primary outcome and first reported outcome differed in 24% of the included RCTs compared with 33% in the included studies other than RCTs. In addition, we found that the proportion of significant P-values is higher in first reported outcomes compared with primary outcomes in RCTs. This is consistent with observations in epidemiological research comparing primary outcomes stated in the protocol with those declared in the final report.[32, 33]

Reporting biases are a prevalent and complex phenomenon across most scientific areas, including epidemiology. Our survey adds to the evidence that statistically significant findings are still overrepresented in current medical research. The phenomenon limits the validity of conclusions drawn from the published literature, and has led to expressions of major concerns and disbelief about the usefulness of scientific evidence.[34, 35] It is important that techniques are used to assess the potential extent of these threats to published evidence, whether in the context of a systematic review or otherwise. Meta-analysis methods provide some of the most direct tools for this, although have major limitations.

Efforts should be increased to tackle the mechanisms causing reporting biases. Initiatives to facilitate registration of studies, protocols and statistical analysis plans are key in this regard. The common practice of interpreting results based on significance tests is likely to have an important role, and researchers should continue to be encouraged to emphasize confidence intervals and effect sizes, rather than P-values, in the interpretation of results.[22, 36, 37] Furthermore, the pressure imposed on researchers to produce scientific publications on a regular basis, coupled with the increasing emphasis on research impact (including journal impact factor), may lead them to dismiss scientific findings for publication if their results are insufficiently innovative or not in agreement with the dominant paradigm. This risks a prioritization of aspects other than rigor and scientific quality when presenting their findings in scientific reports.[38] A new framework in which the importance of replication, as well as the benefits of complete publication of research findings, has been advocated as a promising approach to reduce the prevalence of reporting biases in scientific literature.[25, 39, 40]

Conflict of interest

We declare no conflict of interest.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Ioannidis J. Why most published research findings are false. PLoS Med. 2005; 2: e124.

[2] Parekh-Bhurke S, Kwok CS, Pang C, Hooper L, Loke YK, Ryder JJ, et al. Uptake of methods to deal with publication bias in systematic reviews has increased over time, but there is still much scope for improvement. J Clin Epidemiol. 2011; 64: 349-57.

[3] Rotonda T. Underreporting research is scientific misconduct. JAMA. 1990; 263: 1405-8.

[4] Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database Syst Rev. 2009; 1.

[5] Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, et al. Dissemination and publication of research findings: an updated review of related biases. Health technology assessment. 2010; 14: iii, ix-xi, 1-193.

[6] Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. Bmj. 2011; 343: d4002.

[7] O'Boyle E, Banks G, González-Mulé E. The Chrysalis Effect: How Ugly Initial Results Metamorphosize Into Beautiful Articles. J Manag. 2014; Mar 19: 0149206314527133.

[8] Dwan K, Gamble C, Williamson P, Kirkham J. Systematic review of the empirical evidence of study publication bias and outcome reporting bias—an updated review. PloS ONE. 2013; 8: e66844.

[9] Page M, McKenzie J, Forbes A. Many scenarios exist for selective inclusion and reporting of results in randomized trials and systematic reviews. . J Clin Epidemiol. 2013; 66: 524-37.

[10] Rothstein H, Sutton A, Borenstein M. Publication bias in meta-analysis: Prevention, assessment and adjustments: John Wiley & Sons; 2006.

[11] Sterling T. Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. J Am Stat Assoc. 1959; 54: 30-4.

[12] Rosenthal R. The file drawer problem and tolerance for null results. Psychol Bull. 1979; 86: 638.

[13] van Assen M, van Aert R, Wicherts J. Meta-Analysis Using Effect Size Distributions of Only Statistically Significant Studies. Psychol Methods. 2015; 20: 293.

[14] Simonsohn U, Nelson L, Simmons J. P-curve: A key to the file-drawer. J Exp Psychol Gen. 2014; 143: 534.

[15] Duval S, Tweedie R. Trim and Fill: A Simple Funnel-Plot–Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis. Biometrics. 2000; 56: 455-63.

[16] Song F, Parekh-Bhurke S, Hooper L, Loke YK, Ryder JJ, Sutton AJ, et al. Extent of publication bias in different categories of research cohorts: a meta-analysis of empirical studies. BMC Med Res Methodol. 2009; 9: 1.

[17] Kühberger A, Fritz A, Scherndl T. Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size. PLoS ONE. 2014; 9: e105825.

[18] Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of p-hacking in science. PLoS biology. 2015; 13: e1002106.

[19] Gerber AS, Malhotra N. Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results? Sociological Methods & Research. 2008; 37: 3-30.

[20] Davey J, Turner RM, Clarke MJ, Higgins JP. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. BMC Med Res Methodol. 2011; 11: 160.

[21] Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. Lancet. 2002; 359: 57-61.

[22] Altman DG, Bland JM. How to obtain the P value from a confidence interval. BMJ. 2011; 343: d2304.

[23] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2015.

[24] Bruns SB, Ioannidis JP. p-Curve and p-Hacking in Observational Research. PLoS One. 2016; 11: e0149144.

[25] Ioannidis J, Munafò M, Fusar-Poli P, Nosek B, David S. Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. Trends Cogn Sci. 2014; 18: 235-41.

[26] Palmer A. Quasireplication and the contract of error: lessons from sex ratios, heritabilities and fluctuating asymmetry. Annu Rev Ecol Evol Syst. 2000; 31: 441-80.

[27] Banks GC, Kepes S, Banks K. Publication Bias: The Antagonist of Meta-Analytic Reviews and Effective Policymaking. Educational Evaluation and Policy Analysis. 2012; 34: 259-77.

[28] Gerber A, Green D, Nickerson D. Testing for publication bias in political science. Polit Anal. 2001; 9: 385-92.

[29] Gøtzsche PC. Believability of relative risks and odds ratios in abstracts: cross sectional study. BMJ. 2006; 333: 231-4.

[30] Ioannidis J, Trikalinos T. An exploratory test for an excess of significant findings. Clin Trials. 2007; 4: 245-53.

[31] Masicampo EJ, Lalande DR. A peculiar prevalence of p values just below.05. Q J Exp Psychol. 2012; 65: 2271–9.

[32] Page M, McKenzie J, Kirkham J, Dwan K, Kramer S, Green S, et al. Bias due to selective inclusion and reporting of outcomes and analyses in systematic reviews of randomised trials of healthcare interventions (Review). Cochrane Database Syst Rev. 2014; 10.

[33] Mathieu S, Boutron I, Moher D, Altman D, Ravaud P. Comparison of Registered and Published Primary Outcomes in Randomized Controlled Trials. Jama. 2009; 302: 977-84.

[34] Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nature reviews Drug discovery. 2011; 10: 712.

[35] Sarewitz D. Beware the creeping cracks of bias. Nature. 2012; 485: 149.

[36] Sterne J, Davey Smith G. Sifting the evidence—what's wrong with significance tests? Physical Therapy. 2001; 81: 1464-9.

[37] American P, Association, (APA). The publication manual of the American Psychological Association. 6th ed. Washington, DC: American Psychological Association; 2010.

[38] Fanelli D. Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data. PLoS ONE. 2010; 5: e10271.

[39] Nosek B, Alter G, Banks G, Borsboom D, Bowman S, Breckler S, et al. Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility. Science (New York NY). 2015; 348: 1422.

[40] Van Assen M, Van Aert R, Nuijten M, Wicherts J. Why publishing everything is more effective than selective publishing of statistically significant results. PLoS ONE. 2014; 9: e84896.

Tables

Table 1. Caliper test for reporting biases in first reported outcomes and primary outcomes.

	5% caliper			10% caliper		
	Under caliper	Over caliper	P-value	Under caliper	Over caliper	P-value
First reported outcomes	17	44	<.001	24	60	<.001
Primary outcomes	9	23	0.010	13	35	0.001

Under caliper: number of z-scores that are between 0 and X% smaller than 1.96, with X being the caliper size; **over caliper:** number of z-scores that are between 0 and X% greater than 1.96; P-value: p-value from a one-tailed binomial test

Table 2. Distribution of P-values for all first reported outcomes and primary outcomes.

	P-values >0.05	P-values ≤0.05 & >0.01	P-values ≤0.01
First reported outcomes	166 (21.9%)	216 (28.5%)	376 (49.6%)
Primary outcomes	118 (30.6%)	101 (26.2%)	167 (43.3%)

The distribution of the P-values for the first reported outcomes compared to primary outcomes was performed by the analysis of frequencies ($\chi^2_2 = 10.412$; *p*-value = 0.005).

Table 3. Distribution of P-values for all first reported outcomes and primary outcome grouped by the study design (RCTs vs. studies other than RCTs).

	P-values from RCTs			P-values from studies other than RCTs		
	>0.05	≤0.05 & >0.01	≤0.01	>0.05	≤0.05 & >0.01	≤0.01
First reported outcomes[a]	71 (26.9%)	77 (29.2%)	116 (43.9%)	96 (19.4%)	139 (28.1%)	259 (52.4%)
Primary outcomes[b]	73 (38.6%)	51 (27.0%)	65 (34.4%)	45 (22.8%)	50 (25.4%)	102 (51.8%)

The distribution of the P-values for the first reported outcomes compared to primary outcomes was performed by the analysis of frequencies in included [a] RCTs (χ^2_2 = 7.4666; *p*-value = 0.0239); [b] Studies other than RCTs (χ^2_2 = 1.2052; *p*-value = 0.5474).

Figures

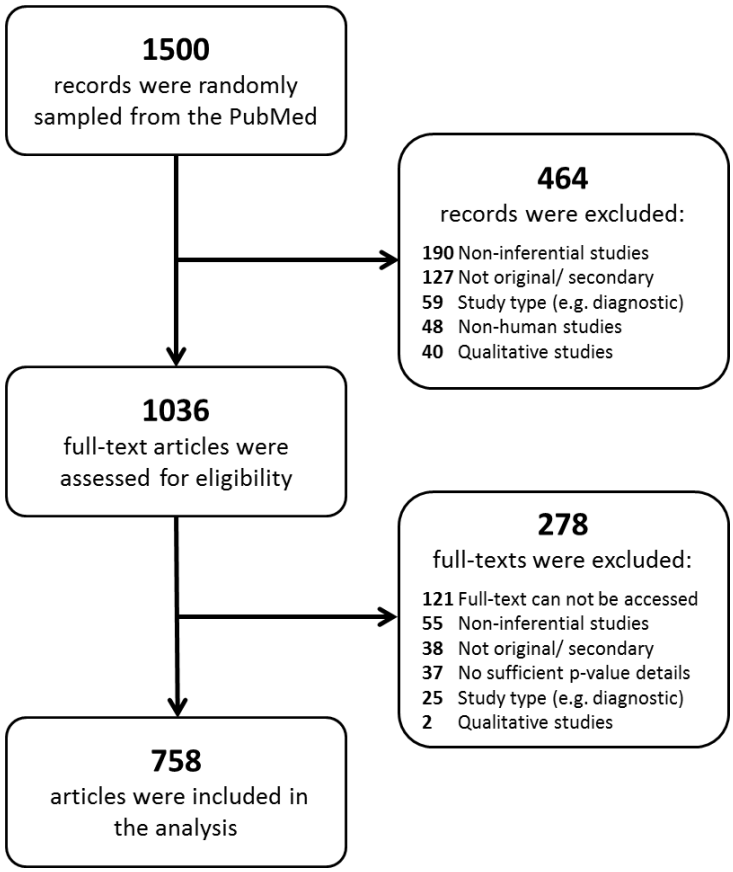


Figure 1. Flow chart for identification of relevant articles from a random sample of records in PubMed

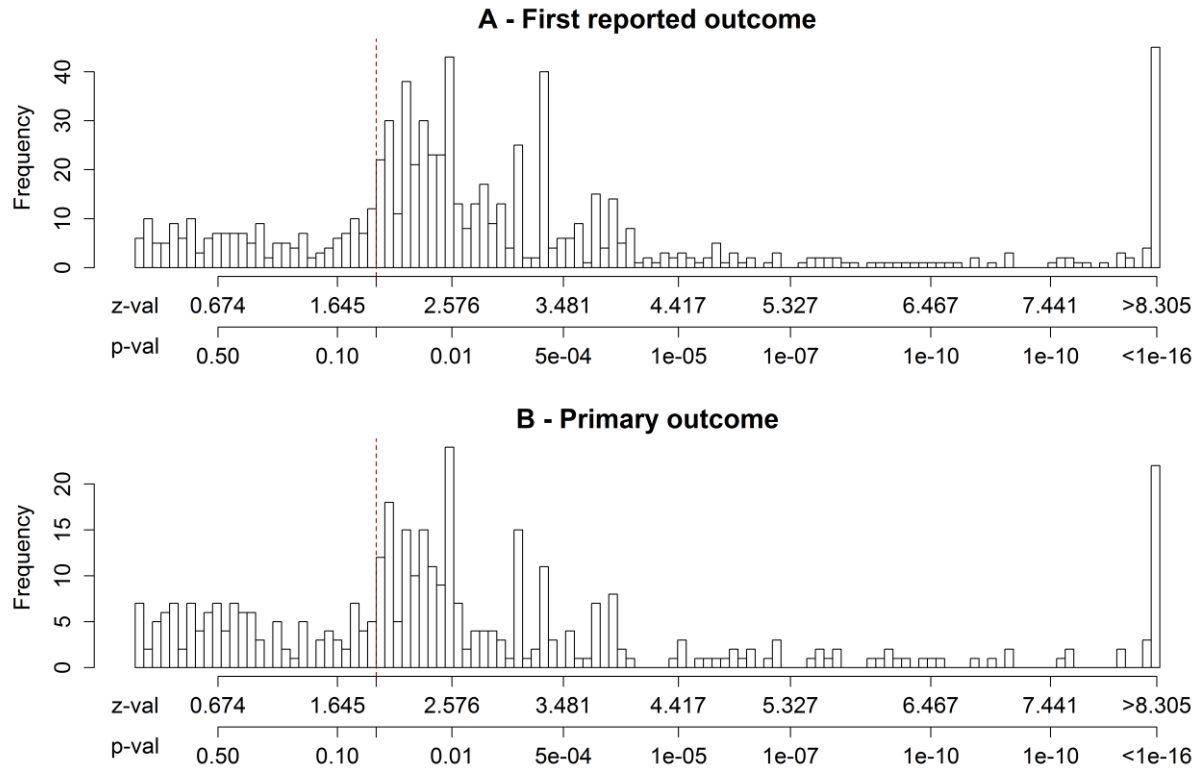


Figure 2. Histograms of z-scores across all articles

These figures display z-scores in absolute value, with the bottom x-axis indicating the corresponding p-value in a two-tailed test; the red dashed line represents the common threshold of p = 0.05 for significance tests.

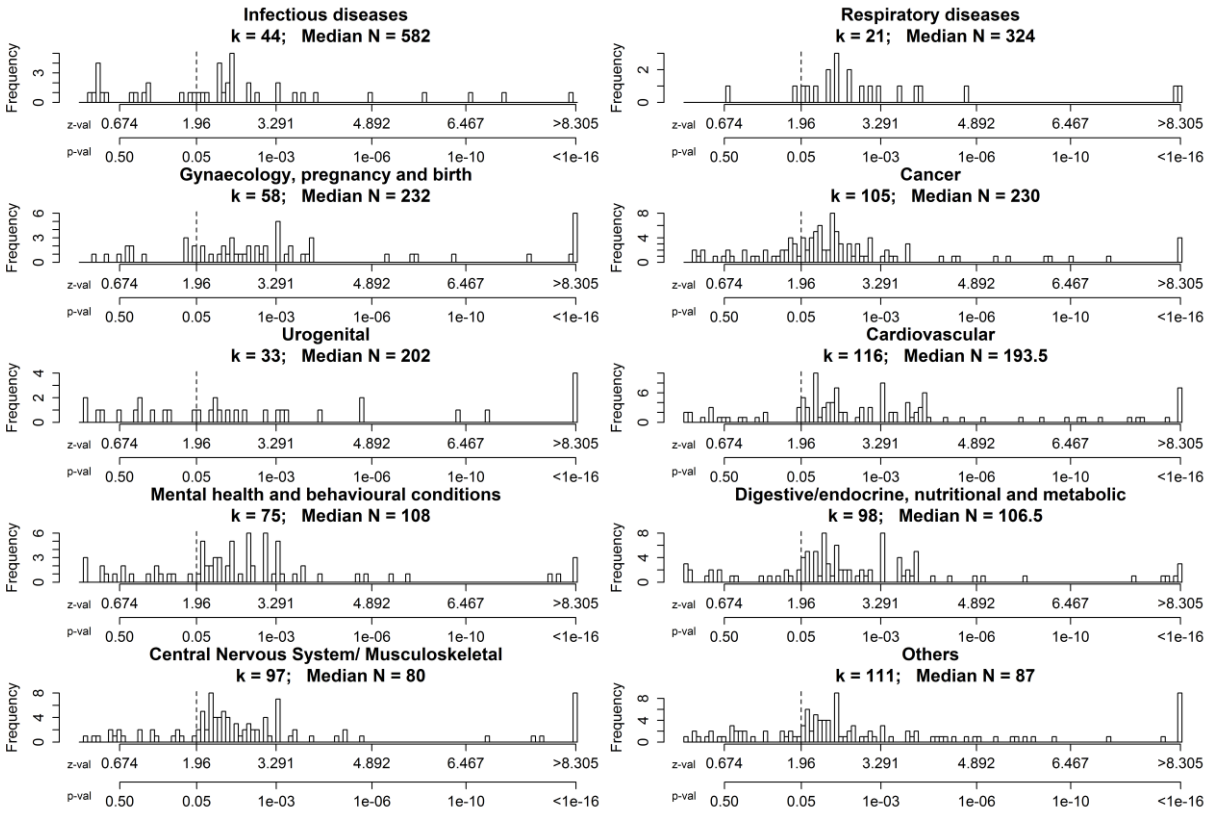
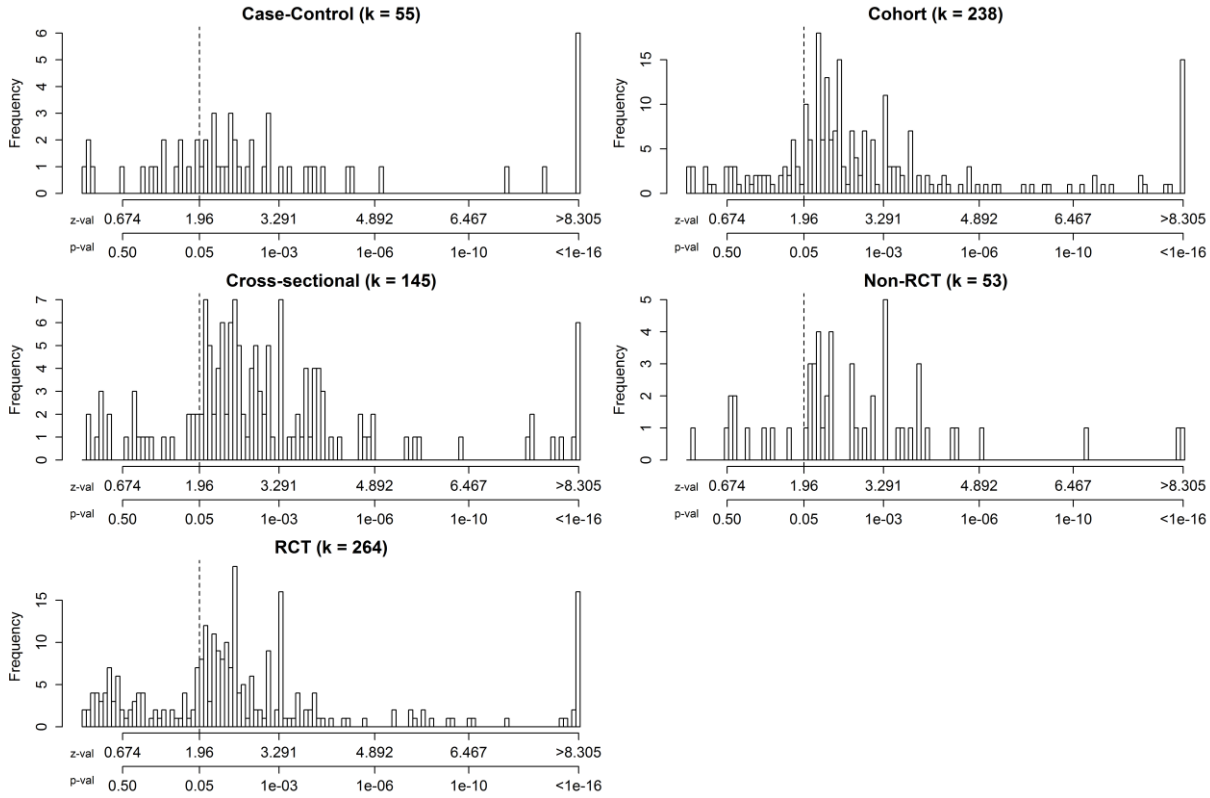


Figure 3. Histograms of z-scores for first reported outcomes, stratified by medical discipline.

These figures display z-scores in absolute value, with the bottom x-axis indicating the corresponding p-value in a two-tailed test; the dashed line represents the common threshold of p = 0.05 for significance tests; k: number of studies; N: sample size.



438

439

440 **Figure 4. Histograms of z-scores for first reported outcomes, stratified by study design**
441 These figures display z-scores in absolute value, with the bottom x-axis indicating the
442 corresponding p-value in a two-tailed test; the dashed line represents the common
443 threshold of p = 0.05 for significance tests; k: number of studies.

444

445

Appendices

Appendix 1. Search strategy for PubMed/Medline

1. "2014/03/01"[Date - Publication] : "2014/03/31"[Date - Publication]
2. Clinical Trial[ptyp]
3. Clinical Trial, Phase III[ptyp]
4. Comparative Study[ptyp]
5. Controlled Clinical Trial[ptyp]
6. Multicenter Study[ptyp]
7. Observational Study[ptyp]
8. Randomized Controlled Trial[ptyp]
9. Pragmatic Clinical Trial[ptyp]
10. Twin Study[ptyp]
11. 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10
12. Humans[Mesh]
13. English[lang]
14. 1 AND 11 AND 12 AND 13

Appendix 2. Supplementary figures

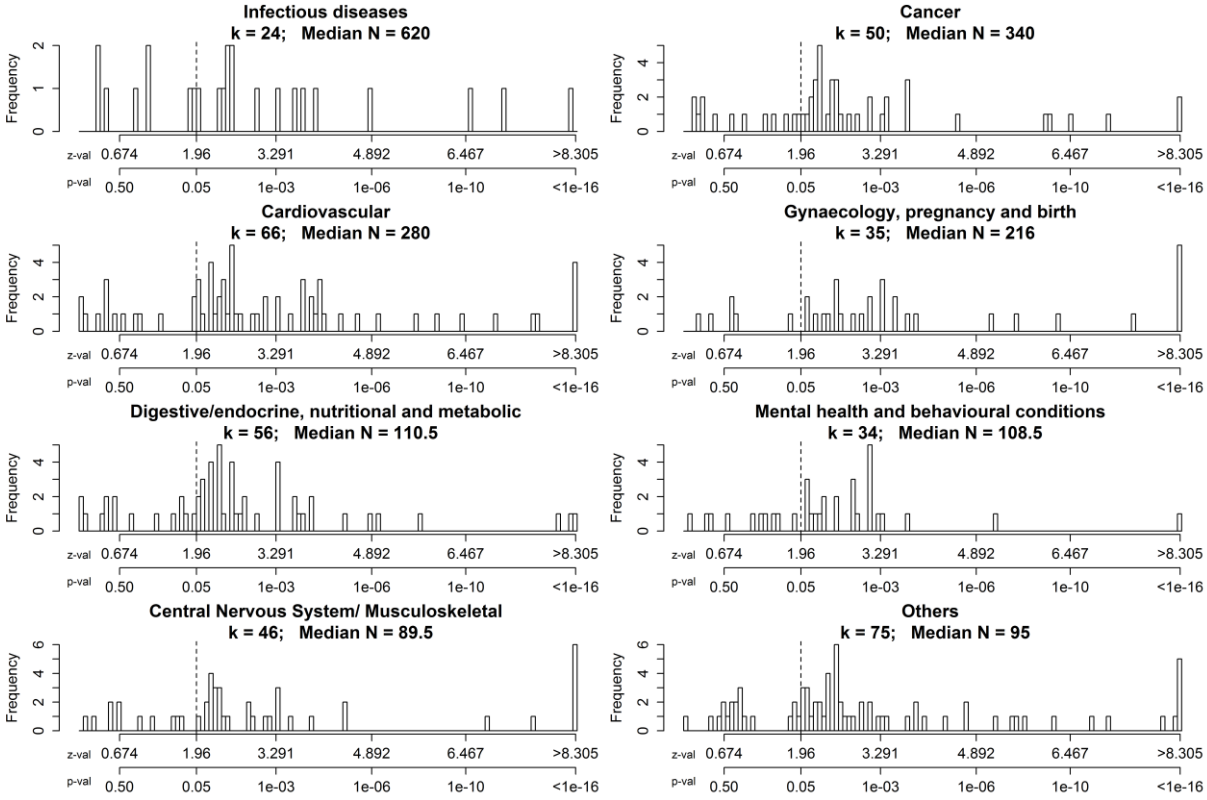


Figure S1. Histograms of z-scores for primary outcomes, stratified by medical discipline

These figures display z-scores in absolute value, with the bottom x-axis indicating the corresponding p-value in a two-tailed test; the dashed line represents the common threshold of p = 0.05 for significance tests; k: number of studies; N: sample size.

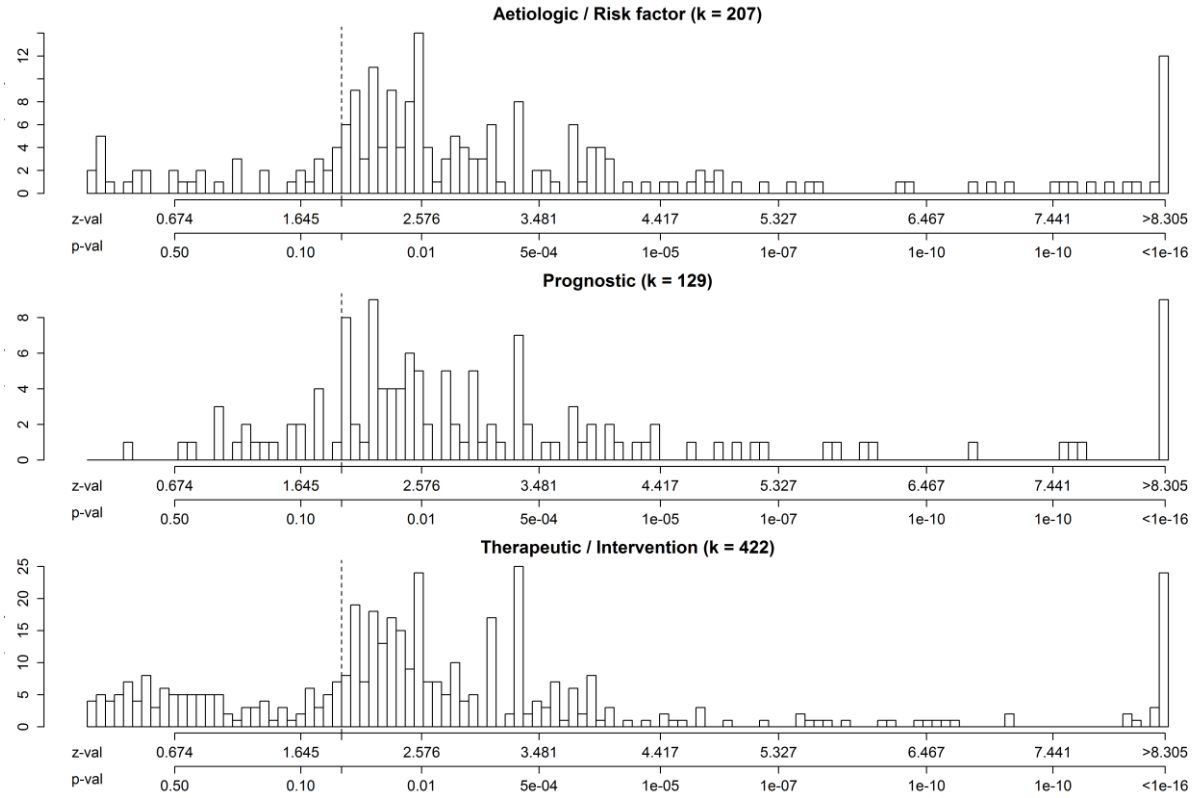


Figure S2. Histograms of z-scores for first reported outcomes, stratified by study type
These figures display z-scores in absolute value, with the bottom x-axis indicating the corresponding p-value in a two-tailed test; the dashed line represents the common threshold of p = 0.05 for significance tests; k: number of studies.

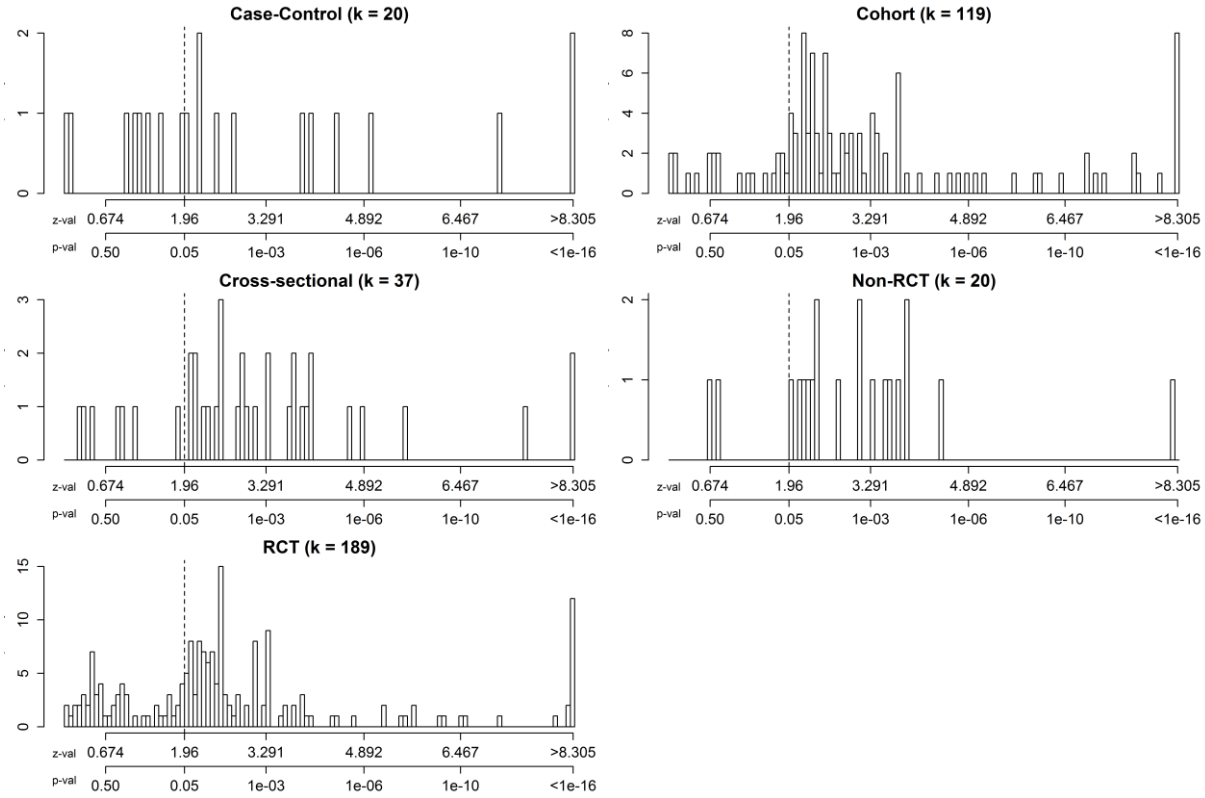


Figure S3. Histograms of z-scores for primary outcomes, stratified by the study design
These figures display z-scores in absolute value, with the bottom x-axis indicating the corresponding p-value in a two-tailed test; the dashed line represents the common threshold of p = 0.05 for significance tests; k: number of studies.

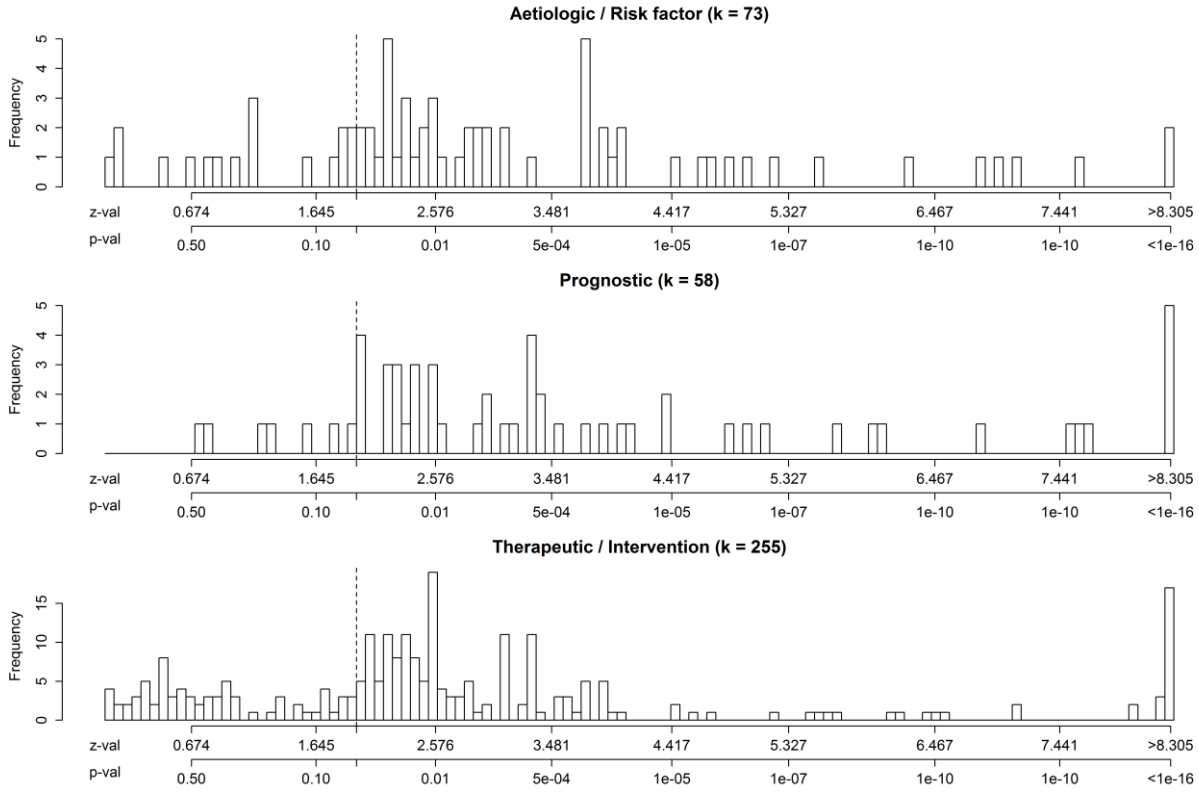


Figure S4. Histograms of z-scores for primary outcomes, stratified by the study type
These figures display z-scores in absolute value, with the bottom x-axis indicating the corresponding p-value in a two-tailed test; the dashed line represents the common threshold of p = 0.05 for significance tests; k: number of studies.